Trabalho: Regressão Linear tendo como variável resposta o IDH do Estado de Santa Catarina

Autor: Marcio Luis de Oliveira Vieira

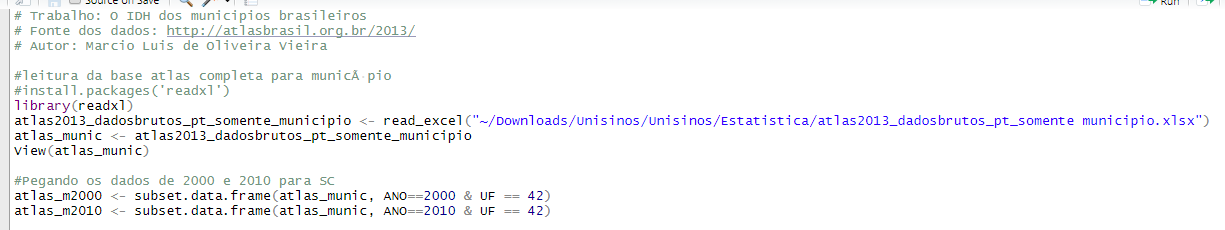
Data: 29/05/2020

**Introdução**

Este trabalho visa demonstrar a criação de um modelo de regressão linear para ter como resposta o IDH de 2010 do Estado de Santa Catarina por município, utilizando como fonte dos dados retirada do Atlas do desenvolvimento Humano no Brasil (<http://www.atlasbrasil.org.br/2013/pt/download>). Todo o trabalho foi desenvolvido utilizando linguagem R e IDE RStudio (<https://rstudio.com/products/rstudio/download/>), todo o código fonte poderá ser encontrado no final deste em anexos.

1. Procedimentos para a coleta:

O Arquivo contendo os dados relativos ao IDH de todos os municípios do Brasil foram baixados do site acima citado para os anos de 2000 e 2010 respectivamente pela professora e disponibilizado em arquivo único para que pudéssemos realizar o trabalho em questão. Com base nestes dados primários, efetuei a leitura do arquivo por completo e depois uma filtragem dos municípios de um determinado estado, conforme sugerido para a tarefa para ambos os anos 2000 e 2010. O Ano de 2000 será utilizado para treinar o modelo de regressão que será posteriormente aplicado ao ano de 2010 para calcular o IDH com base no mesmo.



Nossa amostra para a criação do modelo contém 16695 registros.

1. Procedimentos realizados:

No intuito de escolher as melhores variáveis, executei uma série de testes de correlações e escolhi as 5 melhores variáveis, são elas:

**RDPC** – Renda per capita

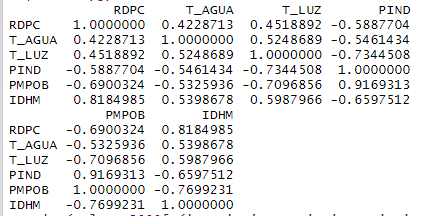
**T\_AGUA** – Percentual da população que vive em domicílios com água encanada

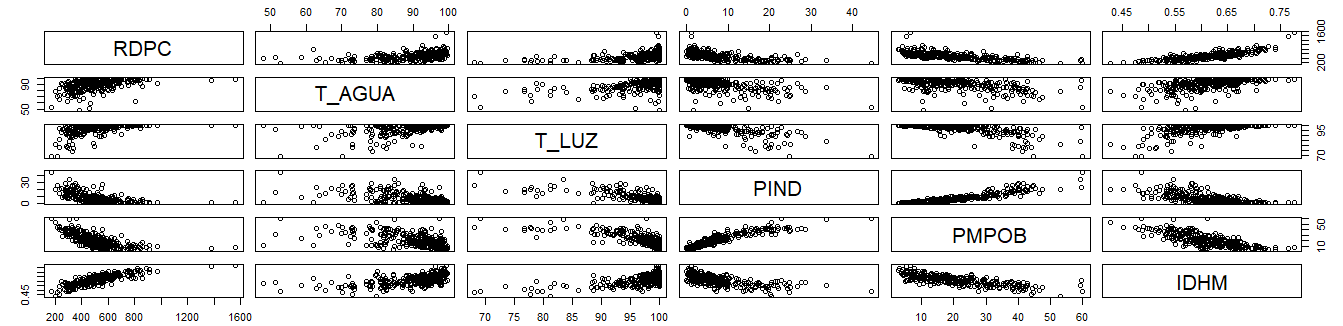
**T\_LUZ** - Percentual da população que vive em domicílios com energia elétrica

**PIND** – Proporção de extremamente pobres

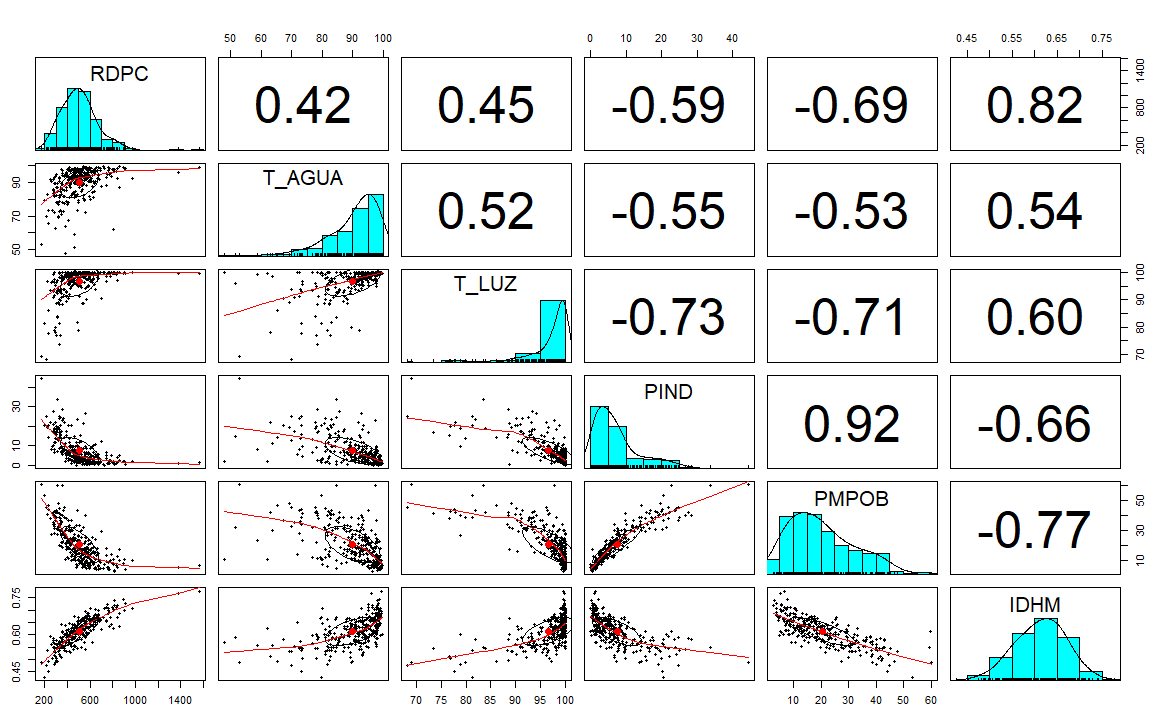
**PMPOB** – Proporção de pobres

Abaixo as análises das correlações destas variáveis independentes com a variável dependente IDH:





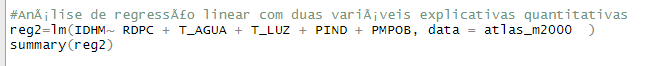
Percebe-se um bom nível de correlação entre as variáveis tanto positiva quanto negativa. Lembrando que a correlação varia entre -1 a 1, quanto mais perto de 1 ou -1, mais forte a correlação positiva ou negativa respectivamente. RPDC por exemplo chegou a 0.8184985.

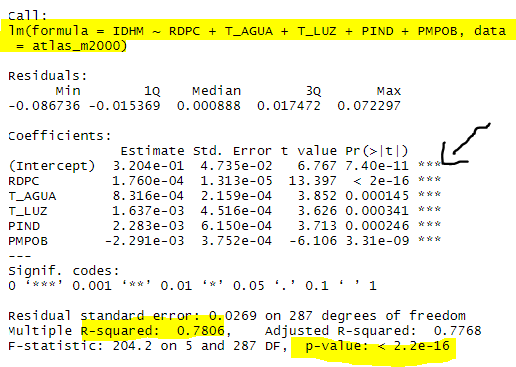


Nesta última observa-se a mesma informação de correlação, mas de uma forma mais rica de informações, inclusive com o histograma de distribuição de cada variável que fará parte da criação do modelo de regressão múltipla.

1. Resultados da modelagem final:

Tendo a variáveis definidas partimos para a criação do modelo de regressão linear utilizando como dados de treino a base de dados de IDH dos municípios do Estado de Santa Catarina para o ano de 2000. E executei comando para verificar e analisar a fórmula e os coeficientes encontrados.

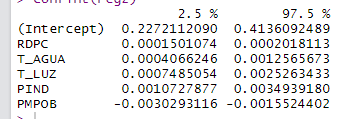




Formalizando, temos a seguinte equação obtida: (formula = IDHM ~ RDPC + T\_AGUA + T\_LUZ + PIND + PMPOB, data = atlas\_m2000).

Como pode ser visto, todas as variáveis apresentam 3 asteriscos, ou seja, nível de significância muito alta para o modelo. Nosso coeficiente de determinação R-squared (R ao quadrado) nos ajuda a avaliar o nível de precisão do modelo, quanto maior, melhor, sendo 1 o valor ideal em nosso caso obtém-se 0.7806. Outro fator importante é o p-value baixo, ou seja, “*uma preditora que tenha um valor-p baixo provavelmente será uma adição significativa ao seu modelo, porque as alterações no valor da preditora estão relacionadas a alterações na variável resposta*”.

Vejamos os intervalos de confiança para os coeficientes da equação utilizando valores padrões da função disponibilizada para o R:

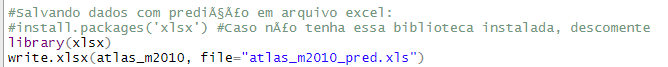


1. Conclusão final:

Com base nisso tudo partimos para a predição usando dados de teste relativos a IDH de 2010 dos municípios do Estado de Santa Catarina para verificarmos a assertividade do modelo:



Na coluna IDHM temos o valor real na coluna fit o valor calculado pelo modelo. Podemos verificar que o modelo possui assertividade satisfatória. As colunas *lwr* and *upr* consideram o desvio padrão. No final gravamos a analise em arquivo excel:



1. Referências:

<https://blog.minitab.com/pt/como-interpretar-os-resultados-da-analise-de-regressao-valores-p-e-coeficientes>

1. Código fonte em R:

# Trabalho: O IDH dos municipios brasileiros

# Fonte dos dados: http://atlasbrasil.org.br/2013/

# Autor: Marcio Luis de Oliveira Vieira

#leitura da base atlas completa para municÃ­pio

#install.packages('readxl')

library(readxl)

atlas2013\_dadosbrutos\_pt\_somente\_municipio <- read\_excel("~/Downloads/Unisinos/Unisinos/Estatistica/atlas2013\_dadosbrutos\_pt\_somente municipio.xlsx")

atlas\_munic <- atlas2013\_dadosbrutos\_pt\_somente\_municipio

View(atlas\_munic)

#Pegando os dados de 2000 e 2010 para SC

atlas\_m2000 <- subset.data.frame(atlas\_munic, ANO==2000 & UF == 42)

atlas\_m2010 <- subset.data.frame(atlas\_munic, ANO==2010 & UF == 42)

#AnÃ¡lise de correlaÃ§Ã£o linear entre duas variÃ¡veis quantitativas

plot(atlas\_m2000$IDHM,atlas\_m2000$RDPC) #Renda per Capita Média

cor(atlas\_m2000$IDHM,atlas\_m2000$RDPC)

plot(atlas\_m2000$IDHM,atlas\_m2000$T\_AGUA) #Percentual com água encanada

cor(atlas\_m2000$IDHM,atlas\_m2000$T\_AGUA)

plot(atlas\_m2000$IDHM,atlas\_m2000$T\_LUZ) #Percentual com Luz elétrica

cor(atlas\_m2000$IDHM,atlas\_m2000$T\_LUZ)

plot(atlas\_m2000$IDHM,atlas\_m2000$PIND) #Proporção extremamente pobre

cor(atlas\_m2000$IDHM,atlas\_m2000$PIND)

plot(atlas\_m2000$IDHM,atlas\_m2000$PMPOB) #Proporção de pobres

cor(atlas\_m2000$IDHM,atlas\_m2000$PMPOB)

#Limpando o dataset para ficar somente com as variaveis importantes para a análise

atlas\_m2000 <- atlas\_m2000[,c('ANO','UF', 'Município','RDPC', 'T\_AGUA', 'T\_LUZ', 'PIND', 'PMPOB', 'IDHM')]

atlas\_m2010 <- atlas\_m2010[,c('ANO','UF', 'Município','RDPC', 'T\_AGUA', 'T\_LUZ', 'PIND', 'PMPOB', 'IDHM')]

class(atlas\_m2000)

#inclusÃ£o do idh2010 na base do ano 2000

idhm\_2010 <- c(atlas\_m2010$IDHM)

class(idhm\_2010)

atlas\_m2000$Y\_idhm\_2010 <- c(atlas\_m2010$IDHM)

str(atlas\_m2000)

#Histograma de algumas variÃ¡veis

hist(atlas\_m2000$RDPC)

hist(atlas\_m2000$T\_AGUA)

hist(atlas\_m2000$T\_LUZ)

hist(atlas\_m2000$PIND)

hist(atlas\_m2000$PMPOB)

summary(atlas\_m2000$IDHM)

hist(atlas\_m2000$IDHM)

cor(atlas\_m2000[c('RDPC', 'T\_AGUA', 'T\_LUZ', 'PIND', 'PMPOB', 'IDHM')])

pairs(atlas\_m2000[c('RDPC', 'T\_AGUA', 'T\_LUZ', 'PIND', 'PMPOB', 'IDHM')])

install.packages('psych')

library(psych)

pairs.panels(atlas\_m2000[c('RDPC', 'T\_AGUA', 'T\_LUZ', 'PIND', 'PMPOB', 'IDHM')])

#AnÃ¡lise de regressÃ£o linear com duas variÃ¡veis explicativas quantitativas

reg2=lm(IDHM~ RDPC + T\_AGUA + T\_LUZ + PIND + PMPOB, data = atlas\_m2000 )

summary(reg2)

#intervalos de confianÃ§a para os coeficientes da equaÃ§Ã£o

confint(reg2)

#resÃ­duos

plot(fitted(reg2),residuals(reg2),xlab="Valores Ajustados",ylab="Resíduos")

abline(h=0)

plot(atlas\_m2000$RDPC,residuals(reg2),xlab="RDPC",ylab="Resíduos")

abline(h=0)

plot(atlas\_m2000$T\_AGUA,residuals(reg2),xlab="T\_AGUA",ylab="Resíduos")

abline(h=0)

plot(atlas\_m2000$T\_LUZ,residuals(reg2),xlab="T\_LUZ",ylab="Resíduos")

abline(h=0)

plot(atlas\_m2000$PIND,residuals(reg2),xlab="PIND",ylab="Resíduos")

abline(h=0)

plot(atlas\_m2000$PMPOB,residuals(reg2),xlab="PMPOB",ylab="Resíduos")

abline(h=0)

#avaliaÃ§Ã£o da suposiÃ§Ã£o de normalidade dos erros,

qqnorm(residuals(reg2), ylab="Resíduos",xlab="Quantis teÃ³ricos",main="")

qqline(residuals(reg2))

#prediÃ§Ã£o da base completa

pred <- predict(reg2, atlas\_m2010, interval="prediction", level=0.95)

atlas\_m2010 <- data.frame(atlas\_m2010,pred)

#Salvando dados com prediÃ§Ã£o em arquivo excel:

#install.packages('xlsx') #Caso nÃ£o tenha essa biblioteca instalada, descomente

library(xlsx)

write.xlsx(atlas\_m2010, file="atlas\_m2010\_pred.xls")